MCVD: Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation

Year: 2022 | Citations: 364 | Authors: Vikram S. Voleti, Alexia Jolicoeur-Martineau, C. Pal

Abstract

Video prediction is a challenging task. The quality of video frames from current state-of-the-art (SOTA) generative models tends to be poor and generalization beyond the training data is difficult. Furthermore, existing prediction frameworks are typically not capable of simultaneously handling other video-related tasks such as unconditional generation or interpolation. In this work, we devise a general-purpose framework called Masked Conditional Video Diffusion (MCVD) for all of these video synthesis tasks using a probabilistic conditional score-based denoising diffusion model, conditioned on past and/or future frames. We train the model in a manner where we randomly and independently mask all the past frames or all the future frames. This novel but straightforward setup allows us to train a single model that is capable of executing a broad range of video tasks, specifically: future/past prediction -- when only future/past frames are masked; unconditional generation -- when both past and future frames are masked; and interpolation -- when neither past nor future frames are masked. Our experiments show that this approach can generate high-quality frames for diverse types of videos. Our MCVD models are built from simple non-recurrent 2D-convolutional architectures, conditioning on blocks of frames and generating blocks of frames. We generate videos of arbitrary lengths autoregressively in a block-wise manner. Our approach yields SOTA results across standard video prediction and interpolation benchmarks, with computation times for training models measured in 1-12 days using $\le$ 4 GPUs. Project page: https://mask-cond-video-diffusion.github.io ; Code : https://github.com/voletiv/mcvd-pytorch